VC-Dimension

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Andrew Ng (Stanford), Eric Eaton (UPenn), David Sontag (NYU), Carlos Guestrin (CMU), Piyush Rai (Utah), and the many others who made their course materials freely available online.

Roadmap

last time

- We will start by analyzing finite hypothesis spaces ($|\mathcal{H}| < \infty$) with zero training error ($R_n(h) = 0$) $\Rightarrow$ **Haussler's Theorem**

- We will then generalize to finite hypothesis spaces ($|\mathcal{H}| < \infty$) with non-zero training error ($R_n(h) > 0$) $\Rightarrow$ **General PAC Bounds**

today

- We will finally discuss infinite hypothesis spaces ($|\mathcal{H}| = \infty$) $\Rightarrow$ **VC-dimension**

PAC Bounds

Given finite hypothesis space $\mathcal{H}$, dataset $\mathcal{D}$ with n iid samples, and probability of error on one sample > ϵ (where $0 \leq \epsilon \leq 1$), then …

Theorem [Haussler '88]

… for any learned hypothesis h that is consistent with the training data ($R_n(h) = 0$),

$$P(R(h) > \epsilon) \leq |\mathcal{H}| e^{-n\epsilon}$$

Theorem [Generalization Bound for $|\mathcal{H}|$ Hypotheses]

… for any learned hypothesis h,

$$P(R(h) - R_n(h) > \epsilon) \leq |\mathcal{H}| e^{-2n\epsilon^2}$$

Based on slides by Carlos Guestrin and David Sontag

Limitations of PAC Bound

With probability at least $1 - \delta$,

$$R(h) \leq \underbrace{R_n(h)}_{\text{bias}} + \underbrace{\sqrt{\frac{1}{2n}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)}}_{\text{variance}}$$

What happens for infinite hypothesis spaces ($|\mathcal{H}| = \infty$), e.g. $\mathcal{H} =$ {all linear classifiers}?

- PAC bound becomes trivial ("infinite" variance)
- We need another way of measuring $|\mathcal{H}|$

Based on slides by Piyush Rai

VC-Dimension

Learning Goals

- Define shattering
- Define VC-dimension

Vapnik-Chervonenkis (VC) Dimension

Goal

Measure "complexity" of a particular class of models independently of training set

Intuition

We only care about the maximum number of points that can be classified correctly

Example

How many points can a linear boundary classify exactly in 1D?

1 point?

2 points?

3 points?

Shattering

Definition

A set $S = \{x^{(1)}, \ldots, x^{(m)}\}$ of points $x^{(i)} \in \mathcal{X}$ is **shattered** by hypothesis class $\mathcal{H}$ if and only if

- for any set of labels $\{y^{(1)}, \ldots, y^{(m)}\}$,
- there exists some consistent $h \in \mathcal{H}$, i.e. $h(x^{(i)}) = y^{(i)}$ for all $i = 1, \ldots, m$.

(Note that S has no relation to the training set.)

More Examples

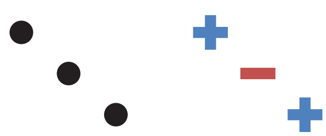
Suppose $\mathcal{H}$ is the set of linear classifiers in 2D.

Can you find a set of 3 points in 2D that $\mathcal{H}$ can shatter?

Based on notes by Andrew Ng

A Note

There may exist a set of 3 points in 2D that $\mathcal{H}$ cannot shatter.



No consistent linear classifier exists for this labeling.

We only care that there exists *at least one* set of 3 points that $\mathcal{H}$ can shatter.

- Rule of thumb: Pick points with maximum separability (e.g. equally spaced along circle).

Continuing our example… Can you find a set of 4 points that $\mathcal{H}$ can shatter? Prove or disprove.

Based on notes by Andrew Ng and example by Eric Eaton

VC-Dimension and Shattering

We use the concept of shattering to define VC-dimension.

To show that hypothesis class $\mathcal{H}$ has VC-dimension d in input space $\mathcal{X}$, consider this adversarial "shattering game":

- We choose d points in $\mathcal{X}$ positioned however we want
- Adversary labels these d points
- We choose a hypothesis $h \in \mathcal{H}$ that separates the points

The VC-dimension of $\mathcal{H}$ in $\mathcal{X}$ is the maximum d we can choose so that we always succeed.

Formal Definition

Given hypothesis class $\mathcal{H}$ and input space $\mathcal{X}$, the **Vapnik-Chervonenkis dimension** $\mathrm{VC}(\mathcal{H})$ over input $\mathcal{X}$ is the size of the largest set of points in $\mathcal{X}$ that is shattered by $\mathcal{H}$.

- If $\mathcal{H}$ can shatter arbitrarily large sets, then $\mathrm{VC}(\mathcal{H}) = \infty$.

Based on notes by Andrew Ng and slides by Piyush Rai

VC-Dimension of Linear Classifiers

For hyperplane with bias, we (informally) showed that…

- VC-dim in $\mathbb{R}^1 = 2$
- VC-dim in $\mathbb{R}^2 = 3$
- VC-dim in $\mathbb{R}^d$?

Recall that such a classifier in $\mathbb{R}^d$ is defined by $d+1$ parameters (one per feature + bias term)

- for linear classifiers, high $d \Rightarrow$ high complexity
- rule of thumb:

Based on slides by Piyush Rai

More VC-Dimension Examples

What is the VC-dimension of 1NN?

What is the VC-dimension of a SVM with RBF kernel?

Based on examples by Piyush Rai [Image from Chris Burges]

Using VC-Dimension in Generalization Bounds

Recall PAC-based generalization bound for hypothesis class $\mathcal{H}$:

$$R(h) \leq R_n(h) + \sqrt{\frac{1}{2n}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)}$$

If $|\mathcal{H}| = \infty$ but $\mathrm{VC}(\mathcal{H}) = d$ in $\mathcal{X}$,

$$R(h) \leq R_n(h) + \sqrt{\frac{1}{2n}\left[d\left(\ln\frac{2n}{d} + 1\right) + \ln\frac{4}{\delta}\right]}$$

where n = training set size
d = VC-dimension of hypothesis class
δ = probability that bound fails

For linear SVM, what does this bound imply?

Note same bias/variance trade-off as always!

Based on slides by Piyush Rai

VC-Dimension of SVMs

But for RBF SVM, $\mathrm{VC}(\mathcal{H}) = \infty$. Is this bad?

- Not really. SVM's large margin property ensures good generalization.

Theorem (Vapnik 1982): Generalization Bound for SVM

- Given n data points $X = \left\{x^{(i)}\right\}_{i=1}^{n}$ such that for all i, $x^{(i)} \in \mathbb{R}^d$ and $||x^{(i)}|| < R$.
- Define $\mathcal{H}_\gamma$ to be the set of classifiers in $\mathbb{R}^d$ with margin γ on X.

Then VC($\mathcal{H}_\gamma$) is bounded by

$$VC(\mathcal{H}_\gamma) \leq \min\left\{d, \left\lceil \frac{4R^2}{\gamma^2} \right\rceil\right\}$$

And with probability $1 - \delta$,

$$R(h) \leq R_n(h) + \sqrt{\frac{1}{2n}\left[VC(\mathcal{H}_\gamma)\left(\ln \frac{2n}{VC(\mathcal{H}_\gamma)} + 1\right) + \ln \frac{4}{\delta}\right]}$$

Note: large $\gamma \Rightarrow$ small VC-dim $\Rightarrow$ low complexity of $\mathcal{H}_\gamma \Rightarrow$ good generalization

Based on slides by Piyush Rai

Learning Theory Take-Aways

- Care about generalization error, not training error
- Standard PAC bounds only apply to finite hypothesis classes
- VC-dimension is measure of complexity of infinite-sized hypothesis classes
- We have formalized the following intuition: suppose we find a model with low training error (low bias)
  - if $|\mathcal{H}|$ large (relative to size of training data), then most likely got lucky (high variance)
  - if $|\mathcal{H}|$ sufficiently constrained and / or large training set, then low training error likely to be evidence of low generalization error (low variance)
- All of this theory is for binary classification

  $\Rightarrow$ it can be generalized to multi-class and regression

Based on slides by Piyush Rai and Eric Eaton